On the robustness of event detection evaluation: a case study

Matthias Feys, Thomas Demeester, Blaz Fortuna, Johannes Deleu, Chris Develder
Ghent University - iMinds, Belgium
firstname.lastname@intec.ugent.be

ABSTRACT

Research on evaluation of IR systems has led to the insight that a robust evaluation strategy requires tests on a large number of events/queries. However, especially for event detection, the number of manually labeled events may be limited. In this paper we investigate how to optimize the evaluation strategy in those cases to maximize robustness. We also introduce two new vector space models for event detection that aim to incorporate bursty information of terms and compare these with existing models. Experiments show that exploiting graded relevance levels reduces the impact of subjectivity and ambiguity of event detection evaluation. We also show that although user disagreement is significant, it has no real impact on result ranking.

Categories and Subject Descriptors

H.2.8 [**Information Systems**]: Database applications - Data mining; H.3.3 [**Information Search and Retrieval**]: Information filtering

Keywords

Event Detection, Evaluation, Vector Space Models

1. INTRODUCTION AND BACKGROUND

Research in the area of event detection deals with discovering events that are described in collections of unstructured texts, such as news data or social media streams, and related tasks such as assigning new content to event streams, or extracting information about entities that play a role in an event[1]. We focus on the subtask of retrospective event detection, where the goal is to discover events in a historical news archive. A popular approach for retrospective event detection is to use the so-called burstiness of terms, i.e., sudden increases in term frequencies [3]. BurstVSM, a method proposed by Zhao *et al.* [5], uses these bursts to define a burst-based Vector Space Model (VSM) and clusters the articles based on this burst-based representation to define the events. We will compare this method with clustering based on tf-idf vectors and two new variants inspired by boosting [2]: increasing the weights of terms that are temporally important.

In contrast to typical evaluation of event detection algorithms, using binary relevance levels and basic metrics such as precision and recall, we will leverage the work of the general IR community [1, 4] and introduce the use of graded relevance levels for this task. We will analyse some robustness issues, ambiguity and subjectivity, of event detection evaluation, and show how graded relevance levels and related metrics can help to reduce the impact of these issues.

[1]E.g., the TAC KBP 2014 event track at `http://www.nist.gov/tac/2014/KBP/Event/index.html`.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.
FIRE 2014 Bangalore, India
Copyright 20XX ACM X-XXXXX-XX-X/XX/XX ...$15.00.

2. EXPERIMENTAL SETUP

2.1 Data Collections

We perform experiments on two separate annotated datasets. The first contains the Chinese articles and annotations used in [5][2], spanning news related web-articles from 2000 to 2009. The second dataset is constructed to test the impact of graded relevance levels and contains Flemish news articles spanning the year 2011[3]. In both datasets the annotated data comprises a list of events (target events) and their corresponding lists of related articles. The latter are constructed by starting from a list of seed events, created by extensive and highly recall-oriented queries, and manually annotating the resulting articles.

2.1.1 Chinese dataset

Since some preprocessing steps could not be reproduced, we were not able to use an exact copy of the set described in [5]. Our version of the dataset contains over 12,000,000 articles and more than 600,000 unique tokens (after filtering out infrequent (df < 10) tokens, and using IKAnalyser[4] for tokenization). The labels per document are binary (a document is either related to the event or not), obtained by using a majority vote of the labels from three different students. In total we have 129 labelled events[5]. As in the original paper, we split the set of events into three partitions based on the number of articles related to each event. The resulting

[2]We like to thank Xin Zhao for sharing his data and code.
[3]Provided by Mediargus (www.mediargus.be)
[4]2012 version of https://code.google.com/p/ik-analyzer/
[5]In [5], only 100 were mentioned.

splits respectively contain 23 *large* (> 300 related articles), 25 *moderate* (containing between 100 and 300 related articles), and 64 *small* events (containing between 10 and 100 related articles). Given that these small events typically do not affect term statistics, and therefore are not interesting for burst-based approaches, we only used the moderate and large events.

2.1.2 Flemish newswire corpus

This set contains around 427,000 Flemish newspaper articles, spanning the year 2011, containing over 98,000 unique tokens (after decapitalization and filtering out words that appear in less than 20 documents or in more than 15% of the documents). We gathered annotations[6] for 19 events, focusing on events with at least 90 more or less relevant articles (corresponding roughly with the moderate and large events of the Chinese dataset, for which the vector space based methods from [5] proved effective). We adopted the following categorical judgements:
Strong (S): The article is strongly and entirely related to the event, either describing it completely, or focusing on a particular aspect.
Weak (W): The article is related to the event, but not necessarily completely, and contains little essential information.
Distant (D): The article is remotely related to the event, but only long before or afterwards (e.g., the event is just mentioned in another context).
No: The article is in no way related to the event.

All 19 events were annotated at least once (spanning almost 12,000 annotated documents in total). For 12 of these events, double judgments were gathered. With U_1 we will refer to the annotators that provided the first judgment for all events, while U_2 provided the second judgment for 12 events. All annotators were experienced and paid students.

2.2 Event-oriented Document Representations

We will compare four vector space models (VSMs) as document representation for event detection: tf-idf, BurstVSM [5], and two new models, their respective boosted counterparts, denoted as B-tf-idf and B-BurstVSM. All methods rely on term features $f(w)$ (i.e., tf-idf) as feature weights.

This boosting is achieved by taking burst information for each term into account. For each term w we identified the corresponding bursts β_i and the burst durations Δ_{β_i} (number of days the term is in a bursty state, q_1). We also calculated the daily emmission costs $\sigma(w, q_i, t)$ for each term and day t, considering the state (bursty q_1 vs. non-bursty q_0) of this term. This emission cost measures the discrepancy between actual and expected term frequency, given the state, as defined in [3]. The boosting for each term is achieved by multiplying the respective static feature weights $f(w)$ with the boosting factor $b^{\beta_i,w,t}$, defined as:

$$b^{\beta_i,w,t} = \log\left[\Delta_{\beta_i} \cdot \max_t\Big(\sigma(q_0,w,t) - \sigma(q_1,w,t)\Big)\right] \quad (1)$$

Where the $\max_t()$ defines the maximum possible gain in daily emmission cost by being in a bursty state (q_1). For all VSMs, the events are subsequently retrieved by grouping the articles based on their respective vector representations.

[6]The annotated Flemish corpus can be made available to researchers upon signing an NDA with Mediargus.

	MAP ± stdev		
	Chinese		Flemish
	moderate	large	all
tf-idf	0.50±0.25	0.35±0.24	0.24±0.22
BurstVSM	0.54±0.21	0.46±0.20	0.22±0.20
B-BurstVSM	*0.59±0.18**	***0.53±0.21****	*0.24±0.20*
B-tf-idf	***0.60±0.21****	*0.53±0.22**	***0.29±0.20***

Table 1: MAP for the different VSMs for the different event sets (see Section 2.1). The best results are in bold, the new VSMs in italic. The systems that significantly outperform tf-idf (p-value <0.05) are indicated with an asterisk.

We used the same clustering tool, CLUTO[7] for all methods. The performance of these methods on the Chinese[8] and Flemish dataset can be found in Table 1. We used Mean Average Precision (MAP) for evaluating the retrieved events (considering only *strong* articles as relevant for the Flemish dataset, and for which the ordering of the articles is based on the distance from the centroid), which enables a natural transition to the graded counterpart, mean graded average precision (GAP). On both datasets we observe that the boosted VSMs outperform their non-boosted counterparts.

Large values for the standard deviation (calculated over all the events) on the metrics are observed on all event sets. We see that on the Chinese dataset, both boosted VSMs significantly outperform the basic tf-idf method, especially for the large events. Note that on the Flemish dataset, the tf-idf method performs better than BurstVSM. Furthermore, none of the observed differences appeared significant at the 0.05 level on this dataset, due to the small number of test events.

3. ANALYSIS OF EVALUATION ISSUES

In this section we take a closer look at the event detection evaluation methodology, focusing on three aspects: (1) the definition of the events, (2) the impact of subjectivity, e.g., by modeling user disagreement, and (3) the impact of the annotation coverage. All experiments are performed on the Flemish dataset.

3.1 Event Definition

To analyse the impact of the event definition on the evaluation results, we consider two aspects. First, we analyse the impact of the relevance definition, i.e., we use binary relevance, but consider different cut-off levels (the relevance categories defined in Section 2.1.2). Secondly, we consider two types of events, homogeneous and heterogeneous, and analyse if the selected event type has an influence on the evaluation results.
Table 2 shows the MAP for three relevance definitions, S: only documents labeled *strong* are considered relevant, $S+W$: both *strong* and *weak* are relevant, and $S+W+D$, including the *distant* documents as well. The results show that the broader $S+W$ definition of event relevance as compared to the S case, leads to improved MAP values for the temporal

[7]www.cs.umn.edu/karypis/cluto, we used the same default settings for standard partitional clustering as in [5]
[8]The difference in effectiveness of the BurstVSM method with respect to [5] is due to the fact that its authors performed a manual preprocessing step in filtering out documents, which was not documented and could not be reconstructed.

	considered relevant		
	S	$S+W$	$S+W+D$
tf-idf	0.24±0.22	0.23±0.21	0.19±0.18
B-tf-idf	**0.29±0.21**	**0.37±0.23***	**0.35±0.22***
BurstVSM	0.22±0.20	0.30±0.23	0.29±0.22
B-BurstVSM	0.24±0.20	0.32±0.23	0.31±0.23

Table 2: MAP values on the Flemish newswire corpus with different relevance cut-off levels. The systems that significantly outperform tf-idf (p-value <0.05) are indicated with an asterisk. Best VSM in bold.

	homogeneous	heterogeneous
tf-idf	**0.30±0.22**	0.19±0.24
B-BurstVSM	0.15±0.13	**0.34±0.23**
BurstVSM	0.12±0.13	0.33±0.21
B-tf-idf	0.26±0.21	0.33±0.22

Table 3: MAP values on the Flemish corpus for two types of events (binary relevance cut-off at the S level).

(burst-based and boosted) methods. The reason is that the temporal VSMs promote key terms with a high temporal value (bursty terms), while ignoring weakly related terms. By only considering the key terms, weakly related articles can be considered more related if they contain the same key terms. By including the *distant* documents, we did not observe a further increase in MAP, rather a decrease, since by definition these documents contain only minimal content related to the event and are published long before or after the event itself and are therefore not retrieved by the algorithms.

The results of the impact of the event types on the system evaluation is shown in Table 3. The homogeneous events are simple events, e.g., articles reporting "the murder of Kadhafi". The heterogeneous events are more complex and consist of multiple types of articles, discussing multiple aspects of the event, e.g., events like "the Occupy Wall Street protest". The decision of event type for each event is based a priori on the description of the event given to the annotators. The results show that the event type can alter the ranking of the methods. Tf-idf is better than the temporal VSMs at detecting homogeneous events and vice versa for the heterogeneous events. Note that Boosted tf-idf performs well for both types. The explanation for these observations is similar to that for the impact of the relevance definition: since the temporal VSMs promote key terms, the impact of the weakly related terms is reduced and heterogeneous articles (containing more weakly related terms) obtain a larger similarity score with these temporal VSMs.

3.2 Subjectivity of the annotations

The user annotations for the evaluation process suffer from subjectivity. This subjectivity can be measured as the (dis)agreement between annotations of different annotators. The Jaccard coefficients between annotations of users U_1 and U_2, when considering different binary relevance definitions are (S): 0.47, (S+W): 0.71, and (S+W+D):0.82.

We observe low overlap if we only consider the articles strongly relevant (S) to the events and a significant increase in overlap if we include the weakly related articles (S+W). The problem is that this more robust relevance cut-off rewards strongly and weakly related articles equally, leading to a lower discriminative power between systems that are better at detecting strongly related articles. We therefore propose to integrate the different relevance levels into the evaluation metric, with the mean Graded Average Precision (shortly denoted as GAP) [4] being the logical extension of MAP. Furthermore, using the so-called user disagreement model [1], we can provide the relevance weights required by GAP, incorporating a probabilistic interpretation by modeling different relevance opinions for each annotation. In Section 3.2.2 we will show GAP can make the evaluation slightly more robust to user disagreement.

	MAP	GAP
U_2 (S)	0.45 ±0.12	0.44±0.10
U_2 ($S+W$)	0.58±0.13	0.63±0.06
U_2 ($S+W+D$)	0.62±0.16	0.73±0.08

Table 4: MAP and GAP with the annotations of U_2 as the retrieved set (binary relevance; relevance definitions as in Table 2). U_1 annotations are used as reference.

3.2.1 User Disagreement Model

The User Disagreement Model (UDM) [1] allows estimating the parameters $P_{T|i}^{M/N}$, the probability that at least M out of a set of N annotators assign a certain document the relevance label T if one observed annotator gave this document the label i. These probabilities are estimated from a small subset of documents annotated by two different annotators, and are subsequently used as relevance weights in the GAP metric. For the calculations of $P_{T|i}^{1/2}$ (based on two annotators), we used micro-averaging over the different events similar to [1]. An important property of the UDM is that it allows evaluation on the highest level of relevance, compensating for those results with a lower relevance that another user might consider as top relevant. This way we can incorporate user disagreement in our evaluation.

3.2.2 Reducing impact of subjectivity with GAP and UDM

We will test if GAP (using the calculated UDM parameters $P_{T|i}^{1/2}$ as the relevance weights, more precisely, 0.21 for the weakly relevant articles, and 0.11 for distant relevant) is more robust to subjectivity (user disagreement) than MAP. We will test this hypothesis with two experiments. First, we will calculate the performance of user U_2, when we consider the annotations of U_1 as reference annotations, using MAP vs. GAP. Next, we will measure the performance of each VSM and calculate the average difference between the performance by using the annotations of user U_1 vs. the annotations of U_2 as reference for both metrics.

The results of the first experiment are shown in Table 4. Since both MAP and GAP require a ranked list of results and the sets of annotations only contain the annotated relevance levels, we calculate both metrics by averaging over 1000 random document orderings within each relevance level. The lower standard deviation on the GAP and the mostly higher absolute values show that using graded relevance levels and the GAP metric leads to a more robust evaluation (w.r.t. user disagreement), especially when event algorithms would output a ranking of all candidate documents in decreasing order of relevance to the event, rather than returning a smaller subset of results for each event, e.g., (S+W(+D)) instead of (S), although the effect is rather limited. Table 5 shows the results of the second experiment. This demonstrates that GAP reduces the mean difference

	$\Delta_{MAP}^{U_1,U_2}$	$\Delta_{GAP}^{U_1,U_2}$
tf-idf	0.07±0.07	0.04±0.04
B-tf-idf	0.16±0.12	0.09±0.07
BurstVSM	0.10±0.09	0.05±0.04
B-BurstVSM	0.13±0.11	0.06±0.06

Table 5: Robustness of evaluation using MAP vs. GAP. The mean and standard deviation of the difference between the results obtained by using the two different reference annotations for all techniques are shown.

between the results obtained by using two different reference annotations. Furthermore, it shows that the temporal VSMs are more susceptible to user disagreement than tf-idf.

3.2.3 *Impact of user disagreement on system rankings*

Finally, we compared the ranking of the different VSMs by using user U_1 vs. U_2 as reference. The system rankings (averaged over all events) are the same for both annotation sets. This shows that the strong user disagreement (cf. Jaccard coefficient of 0.47 when considering only S articles for the comparison of the annotations of U_1 vs. U_2) has little impact on the actual system comparisons. However, by considering the rank correlation of the ranking of the systems for each event individually between the users, averaged over all events, the mean rank correlation increases from 0.53 with MAP, to a mean rank correlation of 0.72 for GAP, again showing the higher robustness of GAP.

3.3 Annotation coverage

In this section we will analyse how many articles of each event we need to annotate in order to make reliable conclusions. Figure 1 shows the mean rank correlation, (and standard deviation) for different levels of annotation coverage using GAP. The rank correlation is defined as the kendall τ's rank correlation between the global ranking (averaged over all events) of systems and the individual system ranking per event, averaged over all events. The mean and standard deviation of the rank correlation for each annotation coverage level are obtained by averaging the rank correlation over 500 random samples (with the given annotation coverage) of the "annotated" articles. The ignored articles are completely removed from the test collection, e.g., not considered relevant, nor non-relevant. The results show that in this experiment, we can almost halve the number of annotations, without altering the resulting rank correlation conclusions[9]. The standard deviation of the rank correlation at this 55% coverage point is 0.05. This means in practice that, for the same annotation cost, twice as many events could be annotated. The standard deviation on the mean evaluation measures would drop accordingly, leading to a better resolution between the systems under comparison.

4. CONCLUSION

We analysed the impact of the ambiguity of the event definition, annotation subjectivity, and annotation coverage on the evaluation of event detection systems. We also introduced two new boosted VSMs that outperform their non-boosted counterparts. Our experiments show that an unambiguous definition of events (specifying the event types

[9]Note that the number of annotations may depend on, e.g., the event type (large vs. small).

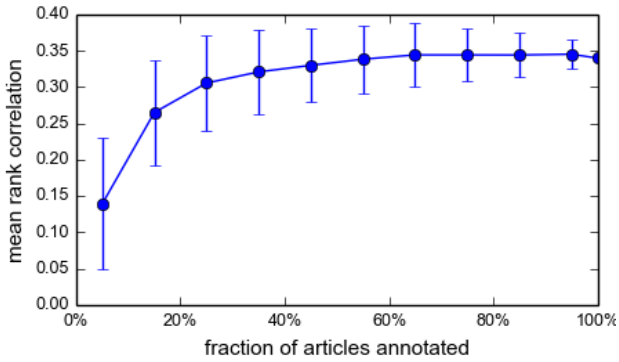

Figure 1: Mean rank correlation for GAP in function of annotation coverage. The mean rank correlations are obtained by sampling 500 times for each annotation coverage point.

and the relevance level between article and event) is critical for comparing between systems. Without precisely defining the required relevance level or choosing a specific type of event, the resulting systems ranking cannot be determined unambiguously. Graded relevance levels can reduce the ambiguity of relevance assessments and hence increase the robustness of the evaluation. The impact of subjectivity of the annotations on the system performance can also be reduced by considering graded relevance levels and using GAP with UDM. Although the user disagreement is large, we showed that it has little impact on the ranking of our systems. This suggests that we do not need to prioritize the problem of user disagreement. However, we need to test this hypothesis on more systems, and in a setting with a larger pool of judged test events, in order to make reliable conclusions. Also, the noticeable influence of the heterogeneity of events needs to be further investigated. Finally, we showed that we can significantly reduce the collection size with little impact on system rankings, opening up the way to larger numbers of annotated test events.

5. ACKNOWLEDGMENTS

This research was carried out in the STEAMER project, facilitated by the iMinds Media Innovation Center (MiX), and financed by the Flemish Agency for Innovation by Science and Technology (IWT).

6. REFERENCES

[1] T. Demeester, R. Aly, D. Hiemstra, D. Nguyen, D. Trieschnigg, and C. Develder. Exploiting user disagreement for web search evaluation: an experimental approach. In *WSDM*, pages 33–42, 2014.

[2] Q. He, K. Chang, E.-P. Lim, and J. Zhang. Bursty feature representation for clustering text streams. In *SDM*, 2007.

[3] J. Kleinberg. Bursty and hierarchical structure in streams. *Data Mining and Knowledge Discovery*, 7(4):373–397, 2003.

[4] S. E. Robertson, E. Kanoulas, and E. Yilmaz. Extending average precision to graded relevance judgments. In *SIGIR*, pages 603–610, 2010.

[5] W. Zhao, R. Chen, K. Fan, H. Yan, and X. Li. A novel burst-based text representation model for scalable event detection. In *ACL*, pages 43–47, 2012.